Bayesian Variable Selection for Linear Regression with the κ-$\mathbf{G}$ Priors

Zichen Ma
Department of Statistics
University of South Carolina
zichen@email.sc.edu

Ernest Fokoué
School of Mathematical Science
Rochester Institute of Technology
epfeqa@rit.edu

Abstract

In this paper, we introduce a new methodology for Bayesian variable selection in linear regression that is independent of the traditional indicator method. A diagonal matrix $\mathbf{G}$ is introduced to the prior of the coefficient vector $\boldsymbol{\beta}$, with each of the g_j's, bounded between 0 and 1, on the diagonal serves as a stabilizer of the corresponding β_j. Mathematically, a promising variable has a g_j value that is close to 0, whereas the value of g_j corresponding to an unpromising variable is close to 1. This property is proven in this paper under orthogonality together with other asymptotic properties. Computationally, the sample path of each g_j is obtained through Metropolis-within-Gibbs sampling method. Also, in this paper we give two simulations to verify the capability of this methodology in variable selection.

Keywords: multiple linear regression; Bayesian variable selection; g-prior

1 Introduction

Consider the traditional multiple linear regression (MLR) model having the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{1.1}$$

where $\mathbf{y}$ is an $n \times 1$ response vector, $\mathbf{X}$ an $n \times p$ data matrix, $\boldsymbol{\beta}$ a $p \times 1$ coefficient vector and $\boldsymbol{\varepsilon}$ the random error. In quite a few areas where the linear model applies, an intreresting yet very important fact is that only a small portion of variables affect the response whereas others are trivial (Jeffreys and Berger, 1991). A great many authors have discussed this topic from both the frequentist (for example, Ullah and Wang (2013)) and the Bayesian perspective (Walli and Wagner, 2011). In this paper, we proceed following the Bayesian path.

In the Bayesian setting (see, for example, Miller (2002) for detail), the coefficient $\boldsymbol{\beta}$ is usually given a conventional g-prior $\mathcal{N}(\mathbf{0}, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$, introduced in Zellner (1986). The g-prior has been

given much attention in Bayesian variable selection primarily because it leads to a computationally tractable Bayes Factor. By introducing an indicator vector, variables are selected and different subsets of variables are compared to each other, or to a reference, based on the value of Bayes Factor. Multiple works have been done to review this methodology. For a recent one, see Dey and Fokoué (2015).

In detail, a random indicator vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_p)^T$ is injected to Equation (1.1), such that for each $\gamma_j, j = 1, 2, \ldots, p$, we have

$$\gamma_j = \begin{cases} 1 & \text{if } \mathbf{x}_j \text{ appears in the model,} \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

Thus, for each combination of γ_j's, Equation (1.1) is modified to

$$\mathbf{y} = \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}),$$

where $\mathbf{X}_{\boldsymbol{\gamma}}$ is the subset of variables according to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the corresponding coefficient vector. There is a total of 2^p combinations of $\boldsymbol{\gamma}$, including the full model, $\boldsymbol{\gamma} = \mathbf{1} = (1, 1, \ldots, 1)^T$, and the null model, $\boldsymbol{\gamma} = \mathbf{0} = (0, 0, \ldots, 0)^T$. For each combination of γ_j's, a corresponding density $p(\mathbf{y} \mid \boldsymbol{\gamma})$ and the Bayes Factor

$$\mathbf{BF}_{\boldsymbol{\gamma}\,\mathbf{1}} = \frac{p(\mathbf{y} \mid \boldsymbol{\gamma})}{p(\mathbf{y} \mid \mathbf{1})}.$$

A difficulty quickly arises when the dimensionality increases, due to the fact that this method searches through the model space of size 2^p. Certain works have been done to solve this problem. George and McCulloch (1993) proposed an empirical method of stochastic search variable selection (SSVS). Each β_j is selected or rejected based on a Monte Carlo average of γ_j, coming from a Gibbs-sampler. Such Monte Carlo average of γ_j is called the posterior inclusion probability (PIP) of β_j. Similar work can be seen in Barbieri and Berger (2004), in which the authors proposed a median probability model rather than a highest probability model, and the variables are selected based on a criterion of $PIP_j > 0.5$. Further, Fokoué (2007) modified the method in Barbieri and Berger (2004) to a prevalence model, which solved the problem that such median probability model may not exist. Certain works have been doen to summarize the Bayesian variable selection with the

indicator method. O'Hara and Sillanpää (2011) provides a thorough review of different methods in Bayesian variable selection. Han and Carlin (2001) gives a comparison in detail of different empirical Bayes methods, especially the Markov Chain Monte Carlo (MCMC) methods, regarding the Bayes Factor.

Certain thoughts have been given to the prior of $\boldsymbol{\beta}$ instead of the traditional g-prior. George and McCulloch (1997) provides a prior of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ that follows

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{D}_{\boldsymbol{\gamma}}\right), \tag{1.3}$$

where $\mathbf{D}_{\boldsymbol{\gamma}}$ is a diagonal matrix and $\mathbf{R}_{\boldsymbol{\gamma}}$ is symmetric. Such prior gives a good generalization of g-prior. Agliari and Parisetti (1988) gives an alternative that follows

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim \mathcal{N}\left(\mathbf{0}, g\sigma^2\left(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{A}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}\right)^{-1}\right), \tag{1.4}$$

where $\mathbf{A}_{\boldsymbol{\gamma}}$ is symmetric and weights different observations, but not the features. Also, see Fernandéz et al. (2001) for a very detailed comparison of different prior choices for Bayesian variable selection. Moreover, multiple works have been done to extend the original Zellner's g-prior. Specifically, Liang et al. (2008) proposed a study on mixtures of g-priors which provides a family of hyperpriors on g while still preserves the tractability on the marginal likelihood. Bové and Held (2011) developed an extension of the classical Zellner's g-prior to generalized linear models, given a large family of hyperpriors on g. Maruyama and George (2011) introduced a fully Bayes formulation with an orthogonal decomposition on the matrix $\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}$, which resolves the issue of $p > n$. All the works mentioned above rely on the indicator method, which is classic but somewhat redundant. To its worst, the methods still have to face the model space of size 2^p. In this work, we intend to get rid of this indicator method completely.

On the other hand, Tipping (2001) introduced a method called the relevance vector machine (RVM) from the machine learning perspective that performs nonparametric variable selection. Retaining the traditional Gaussian prior on $\boldsymbol{\beta}$, with a little modification, each of the β_j's follows a Gaussian prior $(0, \alpha_j^{-1})$ independently. The parameter α_j serves a purpose as the stabilizer. That is, since the coefficient β_j is *a priori* centered at 0, the prior variance become 0 as $\alpha_j \to \infty$, and,

on the contrary, the prior of β_j becomes flat as $\alpha_j \to 0$. Interestingly, as stated in Tipping (2004), combining the non-sparse Gaussian prior on $\boldsymbol{\beta}$ with a Gamma hyperprior on each of the α_j's, the marginal of $\boldsymbol{\beta}$ in fact becomes a multivariate t-distribution after integrating out the α_j's, which leads the RVM to a sparse selection machine. This property of sparsity is even more elegant when the input in the linear model is raised from feature space to kernel space, which is the main focus in Tipping (2001, 2004), but not in our work.

Our work somewhat combines the methodology in George and McCulloch (1997) and Tipping (2004), but gets rid of the traditional indicator method completely. Section 2 provides a thorough theoretical analysis on this new method, including the formulation, some important derivation, and some asymptotic properties. We introduce the computation of model fitting in 3. Here we apply the method of Metropolis-within-Gibbs. In Section 4, we verify the ability of variable selection of this new methodology with two examples. Finally, we provide a summary in Section 5.

2 The κ-G Formulation

2.1 The hierarchical model for variable selection

Given an MLR model with form (1.1), we inject a prior to the coefficient $\boldsymbol{\beta}$ having the form $\mathcal{N}_p(\boldsymbol{\beta} \mid \mathbf{0}, \kappa\sigma^2(\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G})^{-1})$, where, in the variance of the prior, $\kappa > 0$ controls the total scale of the variance, and $\mathbf{G} = diag(g_1, g_2, \ldots, g_p)$ controls how "relevant" each dimension is, with each $g_j \in (0, 1)$ having an impact to the variance of the corresponding β_j. This is to some extent a combination between George and McCulloch (1997) and Tipping (2001). In comparison to George and McCulloch (1997), the diagonal matrix $\mathbf{D}$ in (1.3) is the matrix $\mathbf{G}^{-1}$ here, and $\mathbf{R}$ is the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$. The essential difference is that we have discarded the indicator $\boldsymbol{\gamma}$. Also, in comparison to Tipping (2001), this prior can be seen as a parametric analogy to the prior given in RVM.

Further, each of the g_j's is assigned an i.i.d. $Beta(a, b)$ prior, and by conjugacy κ an inverse-gamma prior $IG(\alpha, \theta)$. We keep the setting in Zellner (1986) for σ^2, that is, a Jeffreys' prior

$p(\sigma^2) \propto (\sigma^2)^{-1}$. And thus, the formulation of the hierarchical model follows:

$$\begin{aligned}
\mathbf{y}\left|\boldsymbol{\beta}, \sigma^2\right. &\sim \mathcal{N}_n\left(\mathbf{y}\left|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right.\right) \\
\boldsymbol{\beta}\left|\mathbf{G}, \kappa, \sigma^2\right. &\sim \mathcal{N}_p\left(\boldsymbol{\beta}\left|\mathbf{0}, \kappa\sigma^2(\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G})^{-1}\right.\right) \\
\mathbf{G} &\sim \prod_{j=1}^{p} Beta\left(g_j\left|a, b\right.\right) \\
\kappa &\sim IG(\alpha, \theta) \\
p(\sigma^2) &\propto (\sigma^2)^{-1}
\end{aligned} \tag{2.1}$$

Directly following (2.1), the joint posterior is given by

$$\begin{aligned}
p\left(\boldsymbol{\beta}, \kappa, \mathbf{G}, \sigma^2\left|\mathbf{y}\right.\right) &\sim p(\mathbf{y}\left|\boldsymbol{\beta}, \sigma^2\right.)p(\boldsymbol{\beta}\left|\mathbf{G}, \kappa, \sigma^2\right.)p(\kappa)p(\mathbf{G})p(\sigma^2) \\
&\sim \mathcal{N}_n\left(\mathbf{y}\left|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right.\right) \times \mathcal{N}_p\left(\boldsymbol{\beta}\left|\mathbf{0}, \kappa\sigma^2(\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G})^{-1}\right.\right) \\
&\quad \times IG(\alpha, \theta) \times \prod_{j=1}^{p} Beta\left(g_j\left|a, b\right.\right) \times (\sigma^2)^{-1} \\
&\sim \left|\sigma^2\mathbf{I}\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right\} \times \\
&\quad \left|\kappa\sigma^2\left(\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\right)^{-1}\right|^{-1/2} \exp\left\{-\frac{1}{2\kappa\sigma^2}\boldsymbol{\beta}^T\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\boldsymbol{\beta}\right\} \\
&\quad \times \kappa^{-\alpha-1} \exp\left(-\frac{\theta}{\kappa}\right) \times \left(\prod_{j=1}^{p} g_j^{a-1}(1-g_j)^{b-1}\right) \times \left(\sigma^2\right)^{-1}.
\end{aligned} \tag{2.2}$$

From (2.2), it is of specific interest to examine the posterior of $\boldsymbol{\beta}$ and $\mathbf{G}$. The former gives some intuition of the connection between this formulation and both the ordinary least square (OLS) estimation and the original Zellner's g-prior, whereas the latter is crucial in the understanding of variable selection with this model.

2.2 Posterior of $\boldsymbol{\beta}$

Following (2.2), the posterior of $\boldsymbol{\beta}$ is given by

$$\begin{aligned} p\left(\boldsymbol{\beta}\left|\kappa, \mathbf{G}, \sigma^2, \mathbf{y}\right.\right) &\sim \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right\} \times \\ &\quad \exp\left\{-\frac{1}{2\kappa\sigma^2}\boldsymbol{\beta}^T\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\boldsymbol{\beta}\right\} \\ &\sim \mathcal{N}_p\left(\tilde{\mu}_{\boldsymbol{\beta}}, \tilde{\Sigma}_{\boldsymbol{\beta}}\right), \end{aligned} \tag{2.3}$$

where $\tilde{\mu}_{\boldsymbol{\beta}}$ and $\tilde{\Sigma}_{\boldsymbol{\beta}}$ are the posterior mean and variance and take one the form of

$$\begin{aligned} \tilde{\mu}_{\boldsymbol{\beta}} &= \left(\mathbf{X}^T\mathbf{X} + \frac{1}{\kappa}\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ \tilde{\Sigma}_{\boldsymbol{\beta}} &= \sigma^2\left(\mathbf{X}^T\mathbf{X} + \frac{1}{\kappa}\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\right)^{-1}. \end{aligned} \tag{2.4}$$

From (2.3) and (2.4), we have the following asymptotic results.

Lemma 2.1. *Denote by $\widehat{\boldsymbol{\beta}}^{(OLS)}$ the OLS estimator of $\boldsymbol{\beta}$. For any $\kappa \neq 0$, as $\mathbf{G} \to \mathbf{0}$, $\tilde{\mu}_{\boldsymbol{\beta}} \to \widehat{\boldsymbol{\beta}}^{(OLS)}$ and $\tilde{\Sigma}_{\boldsymbol{\beta}} \to Var\left(\widehat{\boldsymbol{\beta}}^{(OLS)}\right)$.*

Proof. The proof is rather straightforward. Given $\kappa \neq 0$ and $g_j \to 0$, $\forall j$,

$$\tilde{\mu}_{\boldsymbol{\beta}} \to \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \widehat{\boldsymbol{\beta}}^{(OLS)}$$

and

$$\tilde{\Sigma}_{\boldsymbol{\beta}} \to \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = Var\left(\widehat{\boldsymbol{\beta}}^{(OLS)}\right).$$

□

Lemma 2.2. *For any $\kappa \neq 0$, as $\mathbf{G} \to \mathbf{I}$, we have*

$$\tilde{\mu}_{\boldsymbol{\beta}} \to \frac{\kappa}{\kappa+1}\widehat{\boldsymbol{\beta}}^{(OLS)},$$

which is the same as the posterior mean of $\boldsymbol{\beta}$ in Zellner's g-prior.

Proof. Given $\kappa \neq 0$ and $g_j \to 1$, $\forall j$,

$$\begin{aligned}\tilde{\mu}_{\boldsymbol{\beta}} &\to \left(\mathbf{X}^T\mathbf{X} + \frac{1}{\kappa}\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ &= \frac{\kappa}{\kappa+1}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ &= \frac{\kappa}{\kappa+1}\widehat{\boldsymbol{\beta}}^{(OLS)}\end{aligned}$$

□

Lemma 2.1 states that given $\mathbf{G}$ approaches a null matrix, the posterior mean of $\boldsymbol{\beta}$ approaches the OLS estimator of $\boldsymbol{\beta}$. Also notice that $\mathbf{G} \to \mathbf{0}$ is equivalent to assigning a flat prior to $\boldsymbol{\beta}$, since the prior would have infinite variance. Thus it would lead to a posterior that is equivalent to OLS. Lemma 2.2 states that in the case where $\mathbf{G}$ approaches an identity matrix, the posterior mean of $\boldsymbol{\beta}$ converges to the case in the original Zellner's g-prior, with the parameter κ in this formulation being the same as the original parameter g. This result gives an intuition that the κ-$\mathbf{G}$ formulation is indeed a generalization of Zellner's g-prior. Also, it is of interest that as $\kappa \to \infty$ the convergence from $\tilde{\mu}_{\boldsymbol{\beta}}$ to $\widehat{\boldsymbol{\beta}}^{(OLS)}$ does not require a specific matrix $\mathbf{G}$.

2.3 Posterior of G

We then derive the posterior of $\mathbf{G}$ given $\mathbf{y}$, κ and σ^2 by integrating out $\boldsymbol{\beta}$.

$$\begin{aligned}p\left(\mathbf{G}\,|\mathbf{y},\sigma^2,\kappa\right) &= \int_{\boldsymbol{\beta}} p\left(\mathbf{y}\,|\boldsymbol{\beta},\sigma^2\right) p\left(\boldsymbol{\beta}\,|\mathbf{G}\right) p\left(\mathbf{G}\right) \mathrm{d}\boldsymbol{\beta} \\ &= \prod_{j=1}^{p} Beta\left(g_j\,|a,b\right) \int_{\boldsymbol{\beta}} \mathcal{N}_n\left(\mathbf{y}\,|\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}\right) \times \mathcal{N}_p\left(\boldsymbol{\beta}\,|\mathbf{0},\kappa\sigma^2(\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G})^{-1}\right) \mathrm{d}\boldsymbol{\beta} \\ &\propto |\mathbf{G}|^a\,|\mathbf{I}_p - \mathbf{G}|^{b-1}\left|\mathbf{X}^T\mathbf{X} + \frac{1}{\kappa}\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\right|^{-1/2} \times \\ &\quad \exp\left\{\frac{1}{2}\mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \frac{1}{\kappa}\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\right)^{-1}\mathbf{X}^T\mathbf{y}\right\}\end{aligned} \tag{2.5}$$

Unfortunately, the expression in (2.5) does not have a closed form. However, we could see that the posterior properties of $\mathbf{G}$ relies much on the matrix $\left(\mathbf{X}^T\mathbf{X} + \frac{1}{\kappa}\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\right)^{-1}$. And yet we cannot proceed the analysis of posterior properties of $\mathbf{G}$ in the most general cases since this inverse matrix does not have a further expression in which the matrix $\mathbf{G}$ can be isolated. Figure 1 gives

an intuition of the posterior of $\mathbf{G}$ in the case where $p = 2$. Without loss of generality, we assume $\mathbf{x}_1$ is a promising variable while $\mathbf{x}_2$ is not. In such case, we have $\mathbf{x}_2^T\mathbf{y} = 0$ and $\left|\mathbf{x}_1^T\mathbf{y}\right| \gg 0$. Notice from the figure that the posterior of $\mathbf{G}$ is maximized roughly at $g_1 \to 0$ and $g_2 \to 1$. This is crucial in linking the κ-$\mathbf{G}$ formulation and variable selection. Intuitively, we would expect a promising variable to have a corresponding g_j close to 0 while an unpromising variable to have a g_j close to 1.

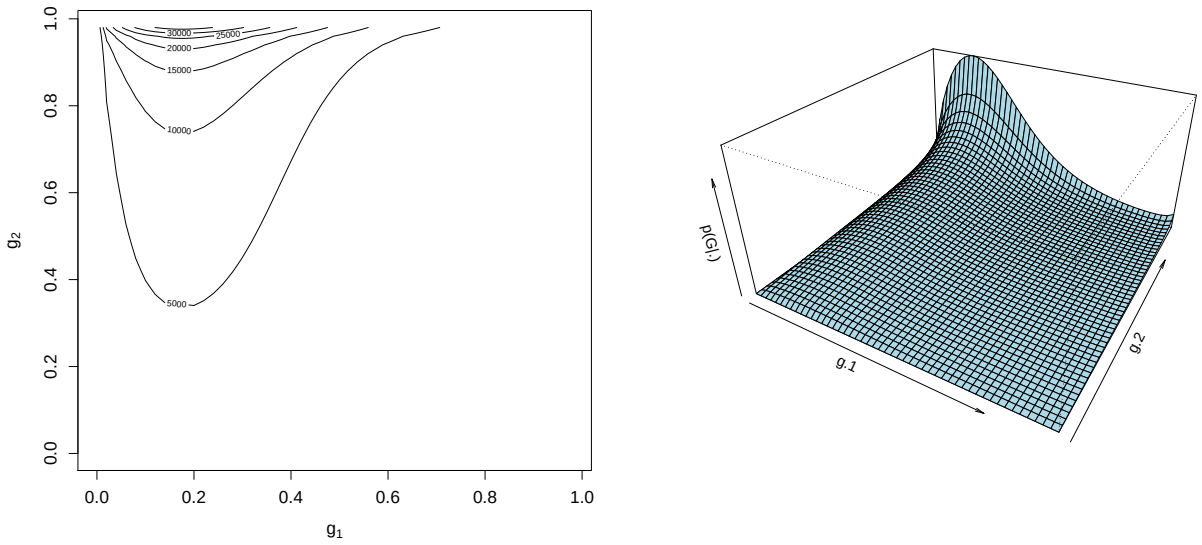

Figure 1: Perspective and contour plot of $p\left(\mathbf{G}\left|\mathbf{y}, \sigma^2, \kappa\right.\right)$

2.4 A case of orthogonality

As stated above, much of the posterior properties rely on $\left(\mathbf{X}^T\mathbf{X} + \frac{1}{\kappa}\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\right)^{-1}$. Though at this point we are not able to proceed to the analysis of the most general case, the analysis under orthogonality where $\mathbf{X}^T\mathbf{X} = diag(\mathbf{x}_j^T\mathbf{x}_j)$ is rather tangible. In this case, the posterior of $\mathbf{G}$ in (2.5) is simplified to

$$p\left(\mathbf{G}\left|\mathbf{y}, \sigma^2, \kappa\right.\right) \propto \prod_{j=1}^{p} \left(g_j^a \left(1 - g_j\right)^{b-1} \left(\kappa + g_j^2\right)^{-1/2} \times \exp\left\{ \frac{\kappa \left(\mathbf{x}_j^T\mathbf{y}\right)^2}{2\sigma^2 \mathbf{x}_j^T\mathbf{x}_j \left(\kappa + g_j^2\right)} \right\} \right). \tag{2.6}$$

Based on (2.6), the joint posterior density of $\mathbf{G}$ can be written as the product of the marginal posterior density functions of each g_j's, which implies that the g_j's are *a posteriori* independent under orthogonality. This simplifies the analysis of $p(\mathbf{G}|\cdot)$ by analysing each individual posterior density $p(g_j|\cdot)$ with

$$p\left(g_j\,|\cdot\right) \propto g_j^a\left(1-g_j\right)^{b-1}\left(\kappa+g_j^2\right)^{-1/2}\exp\left\{\frac{\kappa\left(\mathbf{x}_j^T\mathbf{y}\right)^2}{2\sigma^2\mathbf{x}_j^T\mathbf{x}_j\left(\kappa+g_j^2\right)}\right\}. \tag{2.7}$$

As was mentioned in the introduction, a crucial question with this formulation is:"how is the $\kappa-\mathbf{G}$ methodology linked together with variable selection?" Such question can be seen in two ways. First, we answer how the promising variables lead to certain posterior properties of g_j's. And second, we answer why such properties of g_j's indicate certain variables are promising and others are not.

Theorem 2.1. *A promising variable $\mathbf{x}_j$ has a corresponding g_j that is close to 0, whereas an unpromising variable has a corresponding g_j that is close to 1.*

Proof. Without loss of generality, assume $a=b=\frac{1}{2}$ and $\kappa=\sigma^2=1$. Given the posterior density $p\left(g_j|\mathbf{y},\sigma^2,\kappa\right)$ such that

$$\begin{aligned} p\left(g_j|\mathbf{y},\sigma^2,\kappa\right) &\propto g_j^{1/2}\left(1-g_j\right)^{-1/2}\left(1+g_j^2\right)^{-1/2}\exp\left\{\frac{\left(\mathbf{x}_j^T\mathbf{y}\right)^2}{2\mathbf{x}_j^T\mathbf{x}_j\left(1+g_j^2\right)}\right\} \\ &= g_j^{1/2}\left(1-g_j\right)^{-1/2}\left(1+g_j^2\right)^{-1/2}\exp\left\{\frac{\|\mathbf{y}\|^2\cos^2\theta_j}{2(1+g_j^2)}\right\}, \end{aligned} \tag{2.8}$$

where θ_j is the angle between $\mathbf{x}_j$ and $\mathbf{y}$, the general idea of the proof is that we find the g_j that maximizes the posterior likelihood, i.e. the *maximum a posteriori* estimate for the two cases where $\mathbf{x}_j^T\mathbf{y}=0$ and $\mathbf{x}_j^T\mathbf{y}\neq 0$.

Unpromising variable. For an unpromising variable $\mathbf{x}_j$, it is reasonable to assume that $\cos\theta_j=0$. Therefore in (2.8) $\exp(\cdot)=1$ and we are left with

$$p\left(g_j|\mathbf{y},\sigma^2,\kappa\right) \propto g_j^{1/2}\left(1-g_j\right)^{-1/2}\left(1+g_j^2\right)^{-1/2},$$

which is an increasing function of g_j on $(0,1)$, as $\left(1+g_j^2\right)^{-1/2}$ is monotone decreasing from 1 to $\frac{1}{\sqrt{2}}$, and $g_j^{1/2}(1-g_j)^{-1/2}$ is monotone increasing and $g_j^{1/2}(1-g_j)^{-1/2} \to +\infty$ as $g_j \to 1$. Therefore in the case where the variable $\mathbf{x}_j$ is unpromising we have

$$\widehat{g}_j = \arg\max_{g_j} \; p\left(g_j|\mathbf{y},\sigma^2,\kappa\right) = 1^-. \tag{2.9}$$

Promising variable. For a promising variable $\mathbf{x}_j$, it is reasonable to assume that $\cos\theta_j \approx 1$. Since all the terms on the exponent in (2.8) are positive, $\exp(\cdot)$ is a decreasing function of g_j on $(0,1)$. Further, although the value of $\exp(\cdot)$ somewhat depends on $\|\mathbf{y}\|$, the exponential function dominates the whole posterior likelihood with even a moderate value of $\|\mathbf{y}\|$. Therefore we have

$$\begin{aligned}
\widehat{g}_j &= \arg\max_{g_j} \; p\left(g_j|\mathbf{y},\sigma^2,\kappa\right) \\
&\approx \arg\max_{g_j} \; \exp\left\{\frac{\|\mathbf{y}\|^2\cos^2\theta_j}{2(1+g_j^2)}\right\} \\
&= 0^+.
\end{aligned} \tag{2.10}$$

And thus concludes the proof of the theorem. □

Further, Corollary 2.1 provides a very useful result under orthogonality.

Corollary 2.1. *Under orthogonality, the posterior mean of $\boldsymbol{\beta}^{(Bayes)}$ under the $\kappa-\mathbf{G}$ formulation, $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$, is an unbiased estimator of $\boldsymbol{\beta}$.*

Proof. Denote $\tilde{\mu}_j$ as the posterior mean of the jth variable based on the $\kappa-\mathbf{G}$ formulation. Under orthogonality, that is, $\mathbf{X}^T\mathbf{X} = diag(\mathbf{x}_j^T\mathbf{x}_j)$, the posterior mean of $\boldsymbol{\beta}$ in (2.4) is simplified to

$$\tilde{\mu}_j = \frac{\kappa}{\kappa+g_j}\left(\mathbf{x}_j^T\mathbf{x}_j\right)^{-1}\mathbf{x}_j^T\mathbf{y} = \frac{\kappa}{\kappa+g_j}\widehat{\beta}_j^{(OLS)}.$$

As was shown above, we have $g_j \longrightarrow 0$ for a promising variable. Therefore in this case

$$\tilde{\mu}_j \longrightarrow \frac{\kappa}{\kappa+0}\left(\mathbf{x}_j^T\mathbf{x}_j\right)^{-1}\mathbf{x}_j^T\mathbf{y} = \widehat{\beta}_j^{(OLS)}.$$

Since $\widehat{\beta}_j^{(OLS)}$ is an unbiased estimator of β_j, $\tilde{\mu}_j$ is also unbiased.

On the other hand, if $g_j \longrightarrow 1$, indicating the variable $\mathbf{x}_j$ does not belong to the true model and $\beta_j = 0$, the quantity $\widehat{\beta}_j^{(OLS)}$ should capture the unpromising feature and converges to 0 itself. Therefore the bias also vanishes in this case. □

3 Aspects of Computation

3.1 Conditional density of σ^2 and κ

We then introduce the conditional distribution of κ and σ^2, which mostly serve for the computational purpose. From (2.2), we obtain a closed-form expression of the conditional density of the scale parameter κ,

$$p\left(\kappa \left| \boldsymbol{\beta}, \sigma^2, \mathbf{G}, \mathbf{y}\right.\right) \sim IG(\tilde{\alpha}, \tilde{\theta}), \tag{3.1}$$

where

$$\begin{aligned}\tilde{\alpha} &= \frac{p}{2} + \alpha \\ \tilde{\theta} &= \frac{1}{2\sigma^2}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)^T \mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right) + \theta.\end{aligned}$$

Likewise, the conditional density of σ^2 also has a closed-form expression given by

$$\begin{aligned}p\left(\sigma^2 \left| \boldsymbol{\beta}, \kappa, \mathbf{G}, \mathbf{y}\right.\right) \sim & IG\left(\frac{n+p}{2}\right. \\ & \left.\frac{s^2}{2} + \frac{1}{2}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)^T \mathbf{X}^T\mathbf{X}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) + \frac{1}{2\kappa}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)^T \mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)\right),\end{aligned} \tag{3.2}$$

where

$$\begin{aligned}s^2 &= \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \\ \widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta}^{(OLS)} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.\end{aligned}$$

3.2 A useful sampling algorithm

In this κ-$\mathbf{G}$ formulation, there are four sets of parameters to be estimated from the data. Namely, $\boldsymbol{\beta}$ and $\mathbf{G}$, each consisting of p individual parameters, and κ and σ^2. The MCMC method is very useful in this case to obtain the sample path of the parameters, and specifically, the Gibbs-

sampler is a very convenient tool. However, Gibbs-sampler does require the conditional or posterior density of the parameters to be known, or of closed-form. As we have addressed before, the exact form of the posterior of $\mathbf{G}$ is unknown. Fortunately, the Gibbs sampling of $\mathbf{G}$ can be replaced by a Metropolis step, which only requires the density to be known to a proportion. For each draw of $\mathbf{G}$, the acceptance ratio is

$$r = \frac{p\left(\mathbf{G}^{(*)} \left| \mathbf{y}, \sigma^2, \kappa\right.\right) / J\left(\mathbf{G}^{(*)} \left| \mathbf{G}^{(t-1)}\right.\right)}{p\left(\mathbf{G}^{(t-1)} \left| \mathbf{y}, \sigma^2, \kappa\right.\right) / J\left(\mathbf{G}^{(t-1)} \left| \mathbf{G}^{(*)}\right.\right)} \tag{3.3}$$

where $p(\mathbf{G}|\cdot)$ is given by (2.5) and $J(\cdot)$ is the proposal distribution and is defined as

$$J\left(\mathbf{G}^{(*)} \left| \mathbf{G}^{(t-1)}\right.\right) \sim \prod_{j=1}^{n} Beta\left(g_j^{(*)} |\cdot, \cdot\right).$$

Here we assume that the g_j's within each draw are independent. The shape and scale parameters in $Beta\left(g_j^{(*)} |\cdot, \cdot\right)$ may differ in various cases. As any typical Metropolis-Hastings algorithms, $\mathbf{G}^{(*)}$ is accepted as $\mathbf{G}^{(t)}$ with probability $min(1, r)$. Thus, the whole Metropolis-within-Gibbs algorithm is given in Algorithm 1.

Algorithm 1: The κ-G formulation for Bayesian variable selection

Input: data matrix $\mathbf{X}$, response $\mathbf{y}$, initial values $\boldsymbol{\beta}^{(0)}$, $(\sigma^2)^{(0)}$, $\kappa^{(0)}$ and $\mathbf{G}^{(0)}$

for $t = 1$ **to** T **do**

- Update $\boldsymbol{\beta}^{(t)}$ based on $p\left(\boldsymbol{\beta}^{(t)} \left| \kappa^{(t-1)}, (\sigma^2)^{(t-1)}, \mathbf{G}^{(t-1)}, \mathbf{y}\right.\right)$ as in (2.3);
- Update $\kappa^{(t)}$ based on $p\left(\kappa^{(t)} \left| \boldsymbol{\beta}^{(t)}, (\sigma^2)^{(t-1)}, \mathbf{G}^{(t-1)}, \mathbf{y}\right.\right)$ as in (3.1);
- Update $(\sigma^2)^{(t)}$ based on $p\left((\sigma^2)^{(t)} \left| \boldsymbol{\beta}^{(t)}, \kappa^{(t)}, \mathbf{G}^{(t-1)}, \mathbf{y}\right.\right)$ as in (3.2);
- Accept $\mathbf{G}^{(t)} = \mathbf{G}^{(*)}$ with probability $min(1, r)$ as in (3.3);

end

Notice that the sampling order, that is, which parameters are updated first each time, is mostly arbitrary. We choose to update $\mathbf{G}$ last merely because it involves a Metropolis step, which is more complex than the Gibbs steps.

In terms of varaible selection, we would expect the sample path of g_j's of a promising variable to be severely skewed to the right within in the support of $(0, 1)$, and vice versa. Or in terms of

the posterior mean of g_j, given by

$$\mathbb{E}\left(g_j|\mathbf{y}\right) = \widehat{g_j}^{(Bayes)} = \frac{\sum_{t=1}^{T} g_j^{(t)}}{T}, \tag{3.4}$$

a promising variable would have a $\widehat{g_j}^{(Bayes)}$ that is close to 0, and an unpromising variable close to 1.

4 Numerical Examples and Discussion

4.1 Simulations

In this section we demonstrate our methodology with two simulated examples. First, consider again when $p = 2$. $\mathbf{x}_1$ and $\mathbf{x}_2$ both have 30 observations and come from an i.i.d. $\mathcal{N}(0, 1)$, and the true model is given by

$$y_i = 2x_{i1} + \mathcal{N}(0, 1).$$

Here $\mathbf{x}_1$ is assumed to be the promising variable. Using Algorithm 1, we set the parameters as $a = b = 0.5$, $\alpha = \theta = 1$, and $T = 100000$. In the Metropolis step, we use an independent uniform proposal distribution

$$J\left(\mathbf{G}^{(*)} \middle| \mathbf{G}^{(t-1)}\right) \sim \prod_{j=1}^{n} Beta\left(g_j^{(*)} |1, 1\right).$$

Figure 2 provides a histogram of the sample path of g_j's in the simulation. It is not surprising that g_1 is severely skewed to the right and concentrates toward 0, which corresponds to $\mathbf{x}_1$ being promising, whereas g_2 is severely skewed to the left and concentrates toward 1, corresponding to $\mathbf{x}_2$ being unpromising. Table 1 provides a numerical summary of the g_j's. Due to its severe skewness, here we provide both the mean, denoted by $\widehat{g}_j$, and the median, denoted by $\tilde{g}_j$. The numerical

Table 1: Numerical Summary of g_j, $p = 2$

Variable	$\widehat{\mathbf{g}}_j$	$\tilde{\mathbf{g}}_j$
$\mathbf{x}_1$	**.0682**	**.0427**
$\mathbf{x}_2$	.6986	.8227

summary of g_j for each of the two variable reflects the theoretical deduction in Section 2.

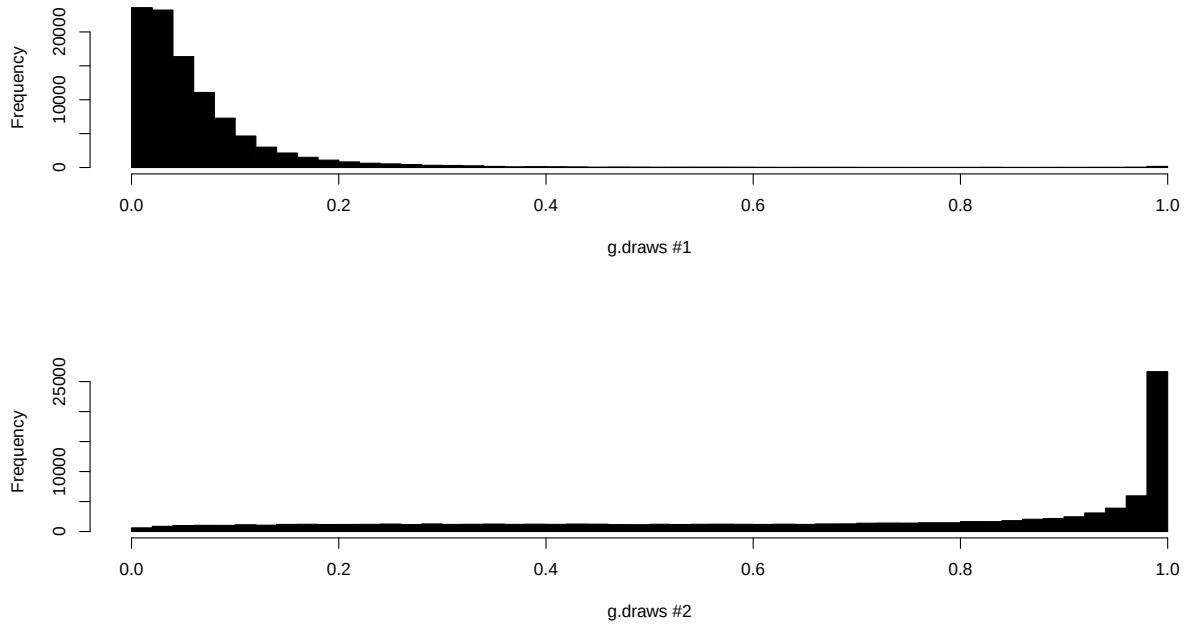

Figure 2: Histogram of g_j, $p = 2$

The second example extends the dimensionality mildly to $p = 10$. Still, all the predictors are i.i.d. from $\mathcal{N}(0, 1)$. The true model is given by

$$y_i = 2(x_{i1} + x_{i2} + x_{i8}) + \mathcal{N}(0, 1).$$

The set-up of the algorithm is mostly the same as in the previous example, except that the prior parameters of g_j are $a = b = 0.3$, instead of 0.5. In this case, the "U" shape of the Beta prior is more strict than before. Also we have $T = 10000$ in this case. Figure 3 provides a comparison of the sample path of the g_j's. Again, we have g_1, g_2, and g_8 close to 0, which corresponds to the associated predictors in the true model.

4.2 Discussion

In Section 1, we introduced how this formulation is motivated by the posterior inclusion probability (PIP) and the relevance vector machine (RVM). Here we discuss these connections in detail using the simulations above.

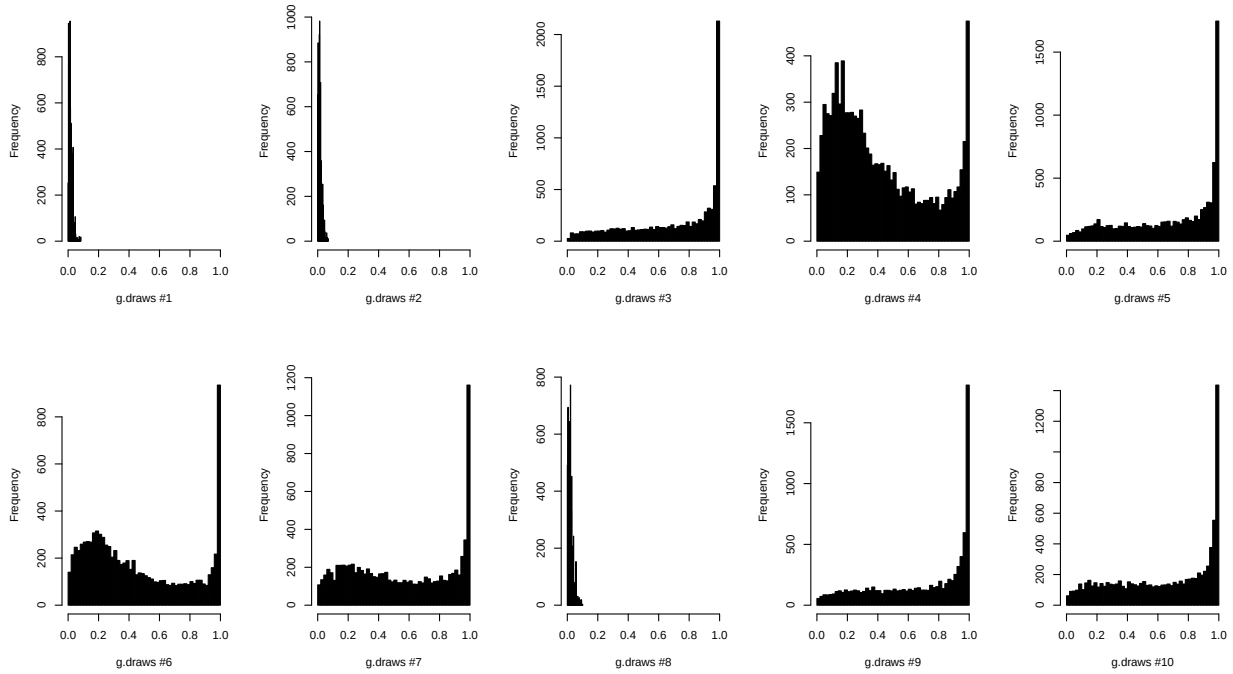

Figure 3: Histogram of g_j, $p = 10$

As stated before, the value $\widehat{g}_j$ or $\tilde{g}_j$ of a promising variable is close to 0, so that the value of $1-\widehat{g}_j$ or $1-\tilde{g}_j$ is close to 1. We can see to this quantity $1-\tilde{g}_j$ as an analogy to the PIP. However, since the procedure of computing PIP searches the space $\gamma_j \in \{0,1\}$, whereas the computation of g_j searches the space $g_j \in (0,1)$, though both quantities are the average of their sample path, quite often the PIP equals to 1 for a promising variable while the value of $\widehat{g}_j$ or $\tilde{g}_j$ can hardly be 0.

Table 2 summarizes the quantities $\widehat{g}_j$, $\tilde{g}_j$, $1-\widehat{g}_j$, $1-\tilde{g}_j$, and the corresponding PIP in the second simulation. For the promising variable $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_8$, g_j's are roughly .01 while the PIPs equal

Table 2: Numerical Summary of g_j, $p = 10$

	$\widehat{\mathbf{g}}_j$	$1-\widehat{\mathbf{g}}_j$	$\tilde{\mathbf{g}}_j$	$1-\tilde{\mathbf{g}}_j$	**PIP**		$\widehat{\mathbf{g}}_j$	$1-\widehat{\mathbf{g}}_j$	$1-\tilde{\mathbf{g}}_j$	$\tilde{\mathbf{g}}_j$	**PIP**
$\mathbf{x}_1$	**.0160**	**.9840**	**.0130**	**.9870**	**1.0000**	$\mathbf{x}_6$	.4517	.5483	.3630	.6370	.0475
$\mathbf{x}_2$	**.0150**	**.9850**	**.0124**	**.9876**	**1.0000**	$\mathbf{x}_7$	.5510	.4490	.5364	.4636	.0403
$\mathbf{x}_3$	.6960	.3040	.7952	.2048	.0386	$\mathbf{x}_8$	**.0196**	**.9804**	**.0157**	**.9843**	**1.0000**
$\mathbf{x}_4$	.4086	.6914	.3186	.6814	.0775	$\mathbf{x}_9$	.6762	.3238	.7713	.2287	.0366
$\mathbf{x}_5$	.6749	.3251	.7652	.2348	.0366	$\mathbf{x}_{10}$	.6352	.3648	.6992	.3008	.0606

to 1, and for the unpromising variables, g_j's are far from 0 while the PIPs are small. Also, it is of interest to notice that the promising variables selected by the two methods are identical although the two methodologies are of different origins.

We then consider the connection between the κ-**G** formulation and the relevance vector machine. One major similarity between the two is the role of the hyper-parameters. Both g_j's in this paper and the α_j's in Tipping (2001) appear in the prior variance of $\boldsymbol{\beta}$. In fact, both g_j and α_j serves as the "stabilizer". That is, given a Gaussian prior centered at 0, a large value of g_j or α_j yields a high prior precision, or low prior variance of β_j, so that the prior of β_j is essentially 0. However, unlike Tipping (2001), in which the prior variance of β_j is solely α_j^{-1}, g_j is only part of the variance, so that it is not necessary to set $g_j \in (0, \infty)$, but only a bounded domain between 0 and 1 is sufficient. Also, in terms of sparsity, the κ-**G** is designed as a sparse machine, that is, we would expect that only a few variables affect the response by assigning a "U-shaped" Beta hyperprior to the parameter g_j.

It is also of interest to verify Corollary 2.1, which indicates, under orthogonality, the unbiasedness of $\widehat{\boldsymbol{\beta}}^{(Bayes)}$ under this formulation. Table 3 provides a comparison of $\widehat{\boldsymbol{\beta}}^{(Bayes)}$ and $\widehat{\boldsymbol{\beta}}^{(OLS)}$ in the second simulation. Given the true values as $\beta_1 = \beta_2 = \beta_8 = 2$, with 10000 iterations, the

Table 3: Comparison of $\widehat{\beta}_j^{(Bayes)}$ and $\widehat{\beta}_j^{(OLS)}$

	κ-**G**	**OLS**		κ-**G**	**OLS**
$\widehat{\beta}_1$	**1.9901**	**1.9939**	$\widehat{\beta}_6$	$-.0306$	$-.0405$
$\widehat{\beta}_2$	**1.9836**	**1.9863**	$\widehat{\beta}_7$	$-.0116$	$-.0192$
$\widehat{\beta}_3$	.0089	.0190	$\widehat{\beta}_8$	**1.9517**	**1.9545**
$\widehat{\beta}_4$	.0543	.0709	$\widehat{\beta}_9$	$-.0071$	$-.0124$
$\widehat{\beta}_5$	.0022	.0023	$\widehat{\beta}_{10}$	.0323	.0551

estimates from the methodology of this paper are very close to the OLS estimates.

5 Conclusion

In this paper we have demonstrated a new methodology for Bayesian variable selection in linear model that is completely independent to the traditional indicator variable method. The coefficient

vector $\boldsymbol{\beta}$ is given a Gaussian prior with the form $\mathcal{N}_p\left(\mathbf{0}, \kappa\sigma^2(\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G})^{-1}\right)$. By injecting a diagonal matrix $\mathbf{G}$ to the variance of the prior, each g_j on the diagonal serves as a variance stabilizer such that the promising variables are selected based on the g_j's that are close to 0. Mathematically, under orthogonality, the g_j's are independent and the posterior of each single g_j is maximized in the support $(0, 1)$ at $g_j \longrightarrow 0$ if the corresponding variable is promising, and vice versa. Further, the estimator of $\boldsymbol{\beta}$ under orthogonality is asymptotically unbiased. Computationally, the hierarchical model is fitted using the Metropolis-within-Gibbs sampling method.

In Section 4, we have demonstrated through two simulations the usefulness of this methodology under orthogonality. Though the dimensionalities in each simulation, $p = 2$ and $p = 10$ respectively, are very mild, the results have shown that this formulation is capable of variable selection and parameter estimation, both with considerable accuracy. The systematic or theoretical examination outside orthogonality is still remained undone, in which the main difficulty involves the inverse matrix $\left(\mathbf{X}^T\mathbf{X} + \frac{1}{\kappa}\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}\right)^{-1}$. In conclusion, as it is completely independent of searching through the 2^p model space, this methodology has the potential of selecting variables with higher efficiency comparing to the traditional methodology and merits further interest and investigation.

References

Agliari, A. and Parisetti, C. (1988). A-g Reference Informative Prior: A Note on Zellner's g-Prior. *Journal of the Royal Statistical Society, Series D*, 37(3):271–275.

Barbieri, M. and Berger, J. (2004). Optimal Predictive Model Selection. *The Annals of Statistics*, 32(3):870–897.

Bové, D. and Held, L. (2011). Hyper g-Priors for Generalized Linear Models. *Bayesian Analysis*, 6(3):387–410.

Dey, T. and Fokoué, E. (2015). Bayesian Variable Selection for Predictive Optimal Regression. In *Current Trends in Bayesian Methodology with Applications*. Chapman and Hall.

Fernandéz, C., Ley, E., and Steel, M. (2001). Benchmark Priors for Bayesian Model Averaging. *Journal of Econometrics*, 100(2):381–427.

Fokoué, E. (2007). Estimation of Atom Prevalence for Optimal Prediction. *Contemporary Mathematics*, 447:103–129.

George, E. and McCulloch, R. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 85:398–409.

George, E. and McCulloch, R. (1997). Approaches for Bayesian Variable Selection. *Statistical Sinica*.

Han, C. and Carlin, B. (2001). Markov Chain Monte Carlo Methods for Computing Bayes Factor: A Comparative Review. *Journal of the American Statistical Association*, 96(455):1122–1132.

Jeffreys, W. and Berger, J. (1991). Sharpening Ockham's Razor on a Bayesian Strop. Technical report, University of Texas at Austin, Purdue University.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). Mixtures of g-Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, 103(481):410–423.

Maruyama, Y. and George, E. (2011). Fully Bayes Factors with a Generalized g-Prior. *The Annals of Statistics*, 39(5):2740–2765.

Miller, A. (2002). *Subset Selection in Regression*. Chapman & Hall/CRC.

O'Hara, R. and Sillanpää, M. (2011). A Review of Bayesian Variable Selection: What, How, and Which. *Bayesian Analysis*, 4(1):85–118.

Tipping, M. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244.

Tipping, M. (2004). Bayesian Inference: An Introduction to Principles and Practice in Machine Learning. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures on Machine Learning*, pages 41–62. Springer.

Ullah, A. and Wang, H. (2013). Parametric and Nonparametric Frequentist Model Selection and Model Averaging. *Econometrics*, 1(2):157–179.

Walli, G. and Wagner, H. (2011). Comparing Spike and Slab Priors for Bayesian Variable Selection. *Austrian Journal of Statistics*, 40(4):241–264.

Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions. In Goel, P. and Zellner, A., editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Amsterdam: North-Holland/Elsevier.